DOCUMENT RESUME

ED 409 334 TM 026 789

AUTHOR Yang, Wen-Ling
TITLE The Effects of Content Mix and Equating Method on the Accuracy of Test Equating Using Anchor-Item Design.
PUB DATE Mar 97
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equated Scores; *Item Response Theory; *Raw Scores; Test Construction; *Test Content; Test Format; Test Items; *True Scores
IDENTIFIERS Accuracy; *Anchor Tests; Anchoring Devices; Three Parameter Model; *Tucker Common Item Equating Method

ABSTRACT

Using an anchor-item design of test equating, the effects of three equating methods (Tucker linear and two three-parameter item-response-theory-based (3PL-IRT) methods), and the content representativeness of anchor items on the accuracy of equating were examined; and an innovative way of evaluating equating accuracy appropriate for the particular item-sampling design of the study was introduced. Data analyzed were test results from 2 forms of a professional competency examination with 197 and 203 items respectively. There were 145 anchor items embedded in both forms, and the 2 examinee groups were not randomly formed. From the two test forms, four pairs of shortened test forms were created to differ in the content representativeness of their anchor items. The total raw score on the original anchor items was regarded as a "pseudo true score," which was used as a criterion for evaluating equating accuracy. Overall, the three equating methods appeared to yield moderately accurate equating results on every test, but the outcomes of the IRT-based methods seemed to be more accurate than the outcomes of the Tucker method. The accuracy of equating depended on the content representativeness of the anchor items, no matter which method was used to equate test forms. The 3PL-IRT model seemed appropriate for equating the test form with negative skewed score distribution. One appendix presents the item sampling schemes and the other contains tables of correlation analyses on anchor and nonanchor items. (Contains 6 tables, 2 figures, and 58 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

Wen-Ling Yang

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 409 334

The Effects of Content Mix and Equating Method on the Accuracy of Test Equating Using Anchor-Item Design

Wen-Ling Yang

Michigan State University

Paper presented at the 1997 AERA Annual Meeting in Chicago

I gratefully acknowledge the valuable contributions of Dr. Richard T. Houang in generating research ideas and advising about statistical analyses.

Inquiries concerning this paper should be sent to Wen-Ling Yang, who is now at the Department of Counseling, Educational Psychology, and Special Education, Michigan State University, East Lansing, MI 48824. (E-mail: yangwenl@pilot.msu.edu)

TM026789

Abstract

Using an anchor-item design of test equating, the effects of three equating methods (Tucker linear and two 3PL-IRT-based methods) and the content representativeness of anchor items on the accuracy of equating were examined in this study. The main goals were to investigate (a) whether equating accuracy improved with more content-representative anchor items, (b) whether the effect of the content representativeness of anchor items depended on the particular equating method used, and (c) relatively, which equating method yielded the most accurate results. An innovative way of evaluating equating accuracy appropriate for the particular item-sampling design of this study was introduced. The adequacy of using the 3PL IRT model for equating alternate forms of a minimum competency test was also discussed.

The data analyzed were test results from two forms of a professional competency examination that had 197 and 203 items respectively. There were 145 anchor items embedded in both forms, and the two examinee groups were not randomly formed. After pooling the two test forms, four pairs of shorter test forms were created by sampling items from the item pool using four distinct item sampling schemes. These item sampling schemes resulted in tests that differed in the content representativeness of their anchor items, and the effect of anchor length was controlled. For each shorter test, the pair of alternate forms were equated using both the conventional linear method and the IRT-based methods.

The total raw score on the 145 anchor items in the original test was regarded as a "pseudo true score", which was used as a criterion for evaluating equating accuracy. Estimated IRT true scores based on the two IRT-based equating and Tucker linear equating result were correlated to "pseudo true score" separately to study the accuracy of these equating. The Pearson produce moment correlation coefficient (r) was used to represent the estimated accuracy of equating results.

In summary, this study found that (a) overall, the three equating methods appeared to yield moderately accurate equating results on every test; (b) however, the equating outcomes of the IRT-based methods seemed to be more accurate than the outcomes of Tucker method, regardless of the content representativeness of anchor items; (c) the two IRT-based methods yielded very similar equating results; (d) the accuracy of equating depended on the content representativeness of anchor items, no matter which method was used to equate test forms; and (e) the 3PL IRT model seemed appropriate for equating the minimum competency test that had negative skewed score distribution.

One important implication of these findings was, regardless of equating method, equating results were more likely to be accurate when anchor items were more representative of the total test, or the content coverage of a test concentrated on fewer topics. Suggestions for future research were provided in this paper.

The Effects of Content Mix and Equating Method on the Accuracy of Test Equating Using Anchor-Item Design

Introduction

In testing practice, often not all examinees take a test at the same occasion or take the same test. To ensure test security, there is a need for alternate test forms. Test forms that have comparable scores are also needed for measuring growth or trends of learning. The need for interchangeable parallel test forms is especially urgent for licensure exams and any other tests used to inform critical decisions. In addition to careful test construction, a practical strategy to arrive at comparable test scores is to establish equivalency between different forms via equating.

A variety of equating techniques have been developed, including linear and non-linear equating. Mainly, equating models vary substantially in their assumptions, mathematical functions, as well as procedures required. Conventional linear methods, such as Tucker linear equating, are straightforward and convenient but their results do not always meet all criteria for equivalent tests. To overcome the drawbacks of conventional equating, equating methods based on IRT estimated scores are developed and used increasingly.

IRT equating is especially useful in common-item design, where random assignment of examinees is not feasible and the assumptions required by conventional equating are likely to be violated (Cook & Eignor, 1991; Crocker & Algina, 1986). Research results have shown that IRT methods are more robust than conventional equating and will lead to greater stability, when tests to be equated differ somewhat in content and length (Petersen, Cook, & Stocking, 1983). Despite various appeals in theory and practice, IRT equating remains under scrutiny because of its sometimes inconsistent behaviors. Possible IRT method by test interaction also raises concerns (Hills, Subhiyah, & Hirsch, 1988; Peterson, Cook, & Stocking, 1983). In addition, practical significance or value of improved accuracy achieved by IRT equating over conventional methods needs to be considered.

To enhance equating accuracy, this study seeks to settle controversies about various equating in practice. Pairs of test forms were assembled by various item sampling schemes to manipulate content mix of a test, or content representation of anchor items embedded in the test. The test forms were then equated by Tucker linear method and two IRT-based equating methods, using anchor-item design. Various equating results were evaluated against an innovative criterion of equating accuracy, which is appropriate for the particular design of this study. Comparisons of equating results are presented and discussed, and suggestions are made for future research and equating practice.

Research Purposes

In search of a better understanding in the function of anchor characteristics in equating and the relative effectiveness of equating methods, this study bears specific purposes as follows:

1. To investigate the effect of content representativeness of anchor items on equating accuracy, while the method of equating varies.
2. To estimate, evaluate, and compare the accuracy of linear equating and IRT-based equating.
3. To compare the equating results of two IRT equating methods (two-stage method and fixed-b method) that are based on different procedures.
4. To apply an innovative criterion for evaluating equating accuracy that is appropriate for the particular design of this study so the effectiveness of various methods can be evaluated.
5. To inform testing practice, based on the findings of this study, about ways to improving equating when anchor-item design is used.

Pursuing solutions to the issues listed above, this study is expected to make contributions to the improvement of test equating practice.

Research Questions

The research questions of this study are shaped by personal interest in understanding and evaluating the effectiveness of various equating methods. They also reflect important equating issues in practice, and they are made viable by the rich context of the data analyzed in this study. To achieve the study goals described previously, the following specific research questions are raised:

1. Does equating result depend on the content representation of anchor items? Specifically, when the content mix of anchor items becomes more representative to the entire test, does the accuracy of equating improve?
2. To what extent do the results of Tucker linear equating and the IRT methods agree, or vary?
3. To what extent do the results of IRT two-stage and IRT fixed-b equating procedures agree, or vary?
4. How accurate are the equating results yielded by various equating methods, compared against an appropriate criterion for evaluating equating accuracy?
5. Is three-parameter logistic (3PL) IRT model appropriate for the minimum competence test, which has a negatively skewed score distribution, analyzed in this study?

Literature Review

Important equating issues, such as conditions of equating, procedures and assumptions of common equating methods, as well as findings from previous research about the merits of various equating methods are reviewed in this section.

Conditions of Equivalency

If test Y is to be equated to test X, no matter what equating procedure is chosen, the following conditions must be satisfied to conclude that the scores on test X and test Y are equivalent (Angoff, 1984; Dorans, 1990; Lord, 1980; Petersen, Kolen, & Hoover, 1989):

1. Both tests measure the same construct.
2. The equating achieves equity. That is, for individuals of identical proficiency, the conditional frequency distributions of scores on the two tests are the same.
3. The equating transformation is symmetric. That is, the equating of Y to X is the inverse of the equating of X to Y.
4. The equating transformation is invariant across sub-groups of the population, from which it is derived.

Equating Guidelines

There is no absolute superior criteria to guide the selection of equating design or method. Arbitrary judgments and decisions that draw on equating expertise and experience are always needed. Factors such as feasibility, cost, and any unique testing context should all be considered.

Brennan and Kolen (1987) argued that the test content and statistical specifications for tests being equated ought to be defined precisely and be stable over time. In the process of test construction, item statistics should be obtained from pre-testing or a previous use of the test. Each test should be reasonably long, with at least 35 items, and the scoring keys should be consistent. The stems for common items, alternatives, and stimulus materials should be identical for the forms to be equated. The characteristics of examinee groups should be stable over time, too. The sizes of the groups should be relatively large, larger than roughly 400. The curriculum, training materials, and field of study should also be stable. The test items should be administered and secured under standardized conditions.

Criteria for Selecting Equating Methods

Usually equating method is selected or tailored to accommodate the need

of a particular testing situation. The three major aspects to be considered for the selection of equating method are reflected in these questions: (1) Are the underlying assumptions required tenable? (2) Is the procedure practical? and (3) How good is the equating result? (Crocker & Algina, 1986)

Tenability of Model Assumptions

The premise of model application is that all the underlying assumptions of the selected model hold. Linear equating assumes that the score distributions of the tests being equated have identical shapes, and is appropriate for equating use when score distributions only differ in the means and/or standard deviations. Equipercentile equating requires fewer assumptions than linear equating. However, in theory, it associates with larger errors than linear equating does (Lord, 1982a). Both linear and equipercentile equating assume that the tests being equated measure the same trait and have equal reliability.

Given tests that have different average difficulty, linear and equipercentile equating are likely to yield erroneous results. The results of these methods also fail to meet the condition of equity and population invariance (Hambleton & Swaminathan, 1990). Unlike these methods, IRT equating does not have the same drawbacks and could be a better alternative.

Applicability of Design and Method

Random groups design, single group design with counter-balancing, and common-item nonequivalent groups design are three common designs used to collect data before equating (Kolen & Brennan, 1995). Random examinee groups design is desirable because each examinee only has to take one form and several forms can be equated at the same time. Nevertheless, it requires the test forms to be available and administered at the same time, which is sometimes not practical. One solution to this problem is the use of anchor design. Either test forms with embedded anchor items (the internal anchor) can be given to different examinee groups, or a third test (the external anchor) can be given to both examinee groups that take different test forms.

Without random assignment, the score distributions of anchor items for different sub-populations may be markedly different and the assumption of equity is unlikely to hold (Crocker & Algina, 1986). In such case, linear or equipercentile method is likely to yield inaccurate results, whereas IRT-based methods seem to have more accurate results.

Equating Accuracy

A major concern for test equating is to what extent the equated scores are equivalent. Random equating errors result from the sampling of examinees and can be controlled by using large examinee samples and choosing appropriate equating designs. Systematic equating errors, whereas, are caused by violations of assumptions and conditions of equating methods. Sometimes, systematic errors can be so large that the results of equating may be worse than no equating (Kolen & Brennan, 1995). To reduce systematic errors, the conditions of equating and assumptions made in equating should be carefully examined.

Perfect equivalency can never be achieved because true score can only be estimated. Consequently, there is no absolute criterion for evaluating equating accuracy. In practice, equating results are often compared against some arbitrary sound criteria to study equating accuracy. Therefore, equating accuracy is an estimate depending on the nature of the arbitrary criterion used. It may be unreasonable to compare all kinds of equating results against one single criterion, because equating methods vary in their assumptions and estimation procedures.

Typically, conventional equating methods that have been known to be satisfactory in yielding accurate results, or have been used in practice for quite a time, are used as evaluation criteria for equating accuracy. Skaggs and Lissitz (1986) argued that the best situation for research purposes was to equate a test to itself through intervening forms.

Tucker Linear Equating

Linear equating has the appeal of simplicity in terms of score transformation and is used most often with the anchor-item design (Kolen &

Brennan, 1987). Among the many linear methods, Tucker linear equating is one of the methods employed most frequently.

Synthetic Population

For anchor-item design, Tucker's method involves the use of a synthetic population (Braun & Holland, 1982). A synthetic population is usually defined as a combination of the proportionally weighted (proportional to sample sizes) populations of examinees taking different test forms. Typically, an equating function is viewed as being defined for a single population, therefore, the two examinee populations must be combined as one single population for defining an equating relationship (Kolen & Brennan,1987).

Model Assumptions

In an anchor-item equating design, suppose Population 1 take Form X , Population 2 take Form Y, and V is the embedded set of anchor items in both forms; to equate scores on Form X to the scale of Form Y, Tucker linear equating requires some strong statistical assumptions as follows (Kolen & Brennan, 1987; Kolen & Brennan, 1995):

1. The linear regression function (slope and intercept) for the regression of X on V is the same for Populations 1 and 2. The function for the regression of Y on V is also the same for the two populations.

2. The variance of X given V is the same for the two populations, and the variance of Y given V is also the same for the two populations.

Under the above assumptions, the linearly transformed scores on one form, yielded by Tucker's method, will have the same mean and standard deviation as the scores on another form. Because of the assumptions about the variances and regression functions in relation to the two populations, Tucker linear equating is more accurate when examinee groups are similar.

Equating Procedures

Using the proportional weights to form a synthetic population, Tucker linear equating basically involves the following concepts and procedures (Kolen & Brennan, 1987; Kolen & Brennan, 1995):

1. Find the weights for Populations 1 and 2 by using these formula: $w_1=n_1/(n_1+n_2)$ and $w_2=n_2/(n_1+n_2)$, where n_1 and n_2 are the sample sizes of examinees from populations 1 and 2 respectively.

2. Let α_1 and α_2 be the regression slopes for the populations, then for Population 1,

$$\alpha_1(X|V)=\sigma_1(X,V)/\sigma_1^2(V) \text{ and } \alpha_1(Y|V)=\sigma_1(Y,V)/\sigma_1^2(V)$$

and for population 2,

$$\alpha_2(X|V)=\sigma_2(X,V)/\sigma_2^2(V) \text{ and } \alpha_2(Y|V)=\sigma_2(Y,V)/\sigma_2^2(V).$$

In addition, let β_1 and β_2 be the regression intercepts for the two populations, and μ_1 and μ_2 be the population means, then

$$\beta_1(X|V)=\mu_1(X)-\alpha_1(X|V)\mu_1(V) \text{ and } \beta_1(Y|V)=\mu_1(Y)-\alpha_1(Y|V)\mu_1(V),$$

and

$$\beta_2(X|V)=\mu_2(X)-\alpha_2(X|V)\mu_2(V) \text{ and } \beta_2(Y|V)=\mu_2(X)-\alpha_2(Y|V)\mu_2(V).$$

To compute the $\hat{\alpha}_1(X|V)$ and $\hat{\alpha}_2(Y|V)$, observed data can be plugged in to the above equations.

3. By assumptions about the slopes and intercepts for the two populations, $\alpha_1(X|V)=\alpha_2(X|V)$, $\alpha_1(Y|V)=\alpha_2(Y|V)$, $\beta_1(X|V)=\beta_2(X|V)$, and $\beta_1(Y|V)=\beta_2(Y|V)$. And, by assumptions about the same variances for the two populations,

$$\sigma_1^2(X)[1-\rho_1^2(X,V)]=\sigma_2^2(X)[1-\rho_2^2(X,V)],$$

and

$$\sigma_1^2(Y)[1-\rho_1^2(Y,V)]=\sigma_2^2(Y)[1-\rho_2^2(Y,V)].$$

4. With the above assumptions, it can be demonstrated that

$$\mu_1(Y)=\mu_2(Y)+\alpha_2(Y|V)[\mu_1(V)-\mu_2(V)],\quad \mu_2(X)=\mu_1(X)-\alpha_2(X|V)[\mu_1(V)-\mu_2(V)],$$

$$\sigma_1^2(Y)=\sigma_2^2(Y)+\alpha_2^2(Y|V)[\sigma_1^2(V)-\sigma_2^2(V)],\quad \sigma_2^2(X)=\sigma_1^2(X)-\alpha_1^2(X|V)[\sigma_1^2(V)-\sigma_2^2(V)],$$

and

$$\sigma_1(Y,V)=\sigma_2(Y,V)[\sigma_1^2(V)/\sigma_2^2(V)],\quad \sigma_2(X,V)=\sigma_1(X,V)[\sigma_2^2(V)/\sigma_1^2(V)].$$

5. The parameters for the synthetic population can be expressed by the weights and the parameters of Populations 1 and 2. The equations for the population means are (a) $\mu_s(X)=w_1\mu_1(X)+w_2\mu_2(X)$, (b) $\mu_s(Y)=w_1\mu_1(Y)+w_2\mu_2(Y)$, and (c) $\mu_s(V)=w_1\mu_1(V)+w_2\mu_2(V)$. And, the population variances are

$$\sigma_s^2(X)=w_1\sigma_1^2(X)+w_2\sigma_2^2(X)+w_1w_2[\mu_1(X)-\mu_2(X)]^2,$$

$$\sigma_s^2(Y)=w_1\sigma_1^2(Y)+w_2\sigma_2^2(Y)+w_1w_2[\mu_1(Y)-\mu_2(Y)]^2,$$

and

$$\sigma_s^2(V)=w_1\sigma_1^2(V)+w_2\sigma_2^2(V)+w_1w_2[\mu_1(V)-\mu_2(V)]^2,$$

where s denotes the synthetic population.

6. Substitute the equations in step 4 in the equations in step 5, the means and variances for the synthetic population on Form X and Form Y can be derived as follows:

$$\mu_s(X)=\mu_1(X)-w_2\alpha_1(X|V)[\mu_1(V)-\mu_2(V)],$$

$$\mu_s(Y)=\mu_2(Y)+w_1\alpha_2(Y|V)[\mu_1(V)-\mu_2(V)],$$

$$\sigma_s^2(X)=\sigma_1^2(X)-w_2\sigma_1^2(X|V)[\sigma_1^2(V)-\sigma_2^2(V)]+w_1w_2\alpha_1^2(X|V)[\mu_1(V)-\mu_2(V)]^2,$$

and

$$\sigma_s^2(Y)=\sigma_2^2(Y)+w_1\sigma_2^2(Y|V)[\sigma_1^2(V)-\sigma_2^2(V)]+w_1w_2\alpha_2^2(Y|V)[\mu_1(V)-\mu_2(V)]^2.$$

To obtain estimates for the means and variances for the synthetic population, plug in observed data to the above equations.

7. After taking the square roots of $\hat{\sigma}_s^2(X)$ and $\hat{\sigma}_s^2(Y)$, the equation for Tucker-linear transformation, $\ell(x)=\sigma_s(Y)/\sigma_s(X)[x-\mu_s(X)]+\mu_s(Y)$, is obtained by replacing the parameters in the above equation with the estimated values obtained previously.

Some Practical concerns

Despite the fact that equal reliability is needed for Tucker linear equating, Kolen and Brennan (1987) argued that, if the test forms were designed to be as similar as possible in content and statistical characteristics, and have the same length, small differences in reliability were not likely to have negative influences on the equating of the two forms.

Compared to Levine equally reliable method, another frequently used linear method that requires the assumption of perfectly correlated (r=1.0) true scores on the two forms, Tucker linear method is often considered more appropriate when examine groups are more similar and test forms less similar. Levine method, whereas, is often said to be more appropriate when test forms are more similar and examinee groups less similar. Nevertheless, research findings have not yet provided clear evidence for the argument (Kolen &

Brennan, 1987).

IRT Equating Methods

Classical methods of equating, developed for equating observed raw scores, are criticized for not being able to meet the conditions of equating (equity, symmetry, and invariance). Equating based on item response theory, whereas, does not suffer from the same drawbacks, given the IRT model fits the data (Hambleton and Swaminathan, 1990; Kolen, 1981). The result of IRT equating, however, varies with the particular equating technique or procedure used. This section provides an overview for IRT equating using anchor-item design.

Linear Transformation of IRT Scales (Two-stage Method)

IRT parameter estimates, obtained from alternate forms of a test, can be converted to the same scale via linear transformation (Kolen and Brennan, 1995). Assuming item and person invariance, linear transformation is reasonable for the non-equivalent-group anchor-item design because the difficulty and discrimination parameters for the common items from the alternate forms are linearly related (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

In theory, given 3PL IRT model fits the data, transformation equations relating IRT parameters for alternate forms of a test (say, Form X and Form Y) are defined as follows (Hambleton and Swaminathan, 1990; Kolen & Brennan, 1995):

(1) For person i, the equation for the ability parameter is $\theta_{y_i} = A\theta_{x_i} + B$, where A and B are constants and θ_{y_i} and θ_{x_i} are the values of person i's ability on the scales of Forms Y and X.

(2) Let a_{y_j}, b_{y_j}, and c_{y_j} be the item parameters for item j on Form Y scale, and a_{x_j}, b_{x_j}, and c_{x_j} be the parameters on Form X scale, (a) the equation for item discrimination parameter is $a_{y_j} = a_{x_j}/A$, (b) the equation for item difficulty parameter is $b_{y_j} = Ab_{x_j} + B$, and (c) the equation for lower asymptote (guessing) parameter is $c_{y_j} = c_{x_j}$.

For a group of persons or items, Kolen & Brennan (1995) showed that the transformation constants (A and B) can be expressed as follows:

$$A = \sigma(b_y)/\sigma(b_x) = \mu(a_x)/\mu(a_y) = \sigma(\theta_y)/\sigma(\theta_x),$$

and

$$B = \mu(b_y) - A\mu(b_x) = \mu(\theta_y) - A\mu(\theta_x).$$

In the above equations, the means $\mu(a_x)$, $\mu(a_y)$, $\mu(b_x)$, and $\mu(b_y)$, as well as the standard deviations $\sigma(b_x)$ and $\sigma(b_y)$, are defined over items. And, the means $\mu(\theta_x)$ and $\mu(\theta_y)$, as well as the standard deviations $\sigma(\theta_x)$ and $\sigma(\theta_y)$, are defined over persons.

In practice, IRT parameters are unknown and thus need to be estimated. In anchor-item equating design, parameter estimates for anchor items can be obtained and used to replace the parameters in the above equations to find the scaling constants. Basically, linear transformation of IRT scales involves two stages: (a) first, alternate test forms are calibrated separately, (b) the information on anchor items obtained from the two IRT calibrations are then used to derive transformation equations for person and item parameters, which can be used to arrive at equivalent scaled scores for examinees taking different test forms.

Other than the above scale-transformation procedure, various techniques

for transforming IRT scales have been proposed. Regression techniques can be applied, but the established relationship is not symmetric (Hambleton and Swaminathan, 1990). The mean/sigma method (Marco, 1977), the mean/mean method (Loyd & Hoover, 1980), and the method involving the use of the geometric means of the a-parameters (Mislevy & Bock, 1990) are all straightforward and similar to the procedure described above. Taking into account individual standard error of estimate, the robust mean and sigma method (Linn, Levine, Hastings and Wardrop, 1981) and robust iterative weighted mean and sigma method (Stocking & Lord, 1983) use variance-weighted means and standard deviations to find the transformation constants. In short, poorly estimated parameters with larger variances receive less weights. The iterative method also weights outliers less.

The above methods, however, suffer from a common flaw; that is, they do not take into account all of the item parameters at the same time. As a result, various combinations of a-, b-, and c-parameter estimates may result in very similar item characteristic curves over the range of the most occurring ability.

Characteristic Curve Transformation (Formula) Methods

Unlike the above methods, characteristic curve methods developed by Haebara (1980) and Stocking and Lord (1983) consider the parameter estimates simultaneously. The two methods estimate the difference between the item characteristic curves on the two scales, for a given θ and over items, differently. However, both methods rely on iterative algorithms that minimize the overall differences over examinees to find the transformation constants (A and B).

It is found from some comparison study that the characteristic curve transformation methods yielded more accurate results than the other methods. Nevertheless, the results did not differ much sometimes (Baker & Al-Karni, 1991). In addition to the computationally intensive iteration procedures, the characteristic curve methods also have the limitation of not explicitly accounting for the error in estimating item parameters (Kolen & Brennan, 1995).

Fixed-b Method

The fixed-b IRT equating method sequentially calibrates test items following these steps:

(1) Estimate bs and other item parameters for Book-A items;

(2) Calibrate Book-B items by fixing bs for the anchor items at the values obtained from the previous step;

(3) Book-B scale is then fixed onto the scale of Book A (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

IRT True-Score Equating

In theory, true scores on alternate tests or test forms can be obtained and equated. To eliminated negative scores and to provide a readily interpretable scale, values on θ (ability) scale may be transformed to their corresponding true score values (Hambleton, Swaminathan, and Rogers, 1991). Then, the true scores on alternate forms can be equated via some linear transformation.

IRT true Scores.

Let θ be the parameter of ability and n be the number of items in a test, true score can be defined as follows: True score $(\xi) = \sum_{1}^{n} p_i(\theta)$ (Crocker and Algina, 1986; Lord, 1980; Hambleton & Swaminathan, 1990). When comparing tests or test forms of different lengths, instead of ξ, true proportion correct or domain score (π) can be reported. Ranging between 0 and 1, π is computed by dividing ξ by the number of items (n) in test forms-- $\pi=\xi/n$ (Hambleton & Swaminathan, 1990; Hambleton, Swaminathan, and Rogers, 1991).

Taking into account the numbers of alternative options, which has substantial influence on guessing, the true score formula can be rewritten to (Petersen, Cook, & Stocking, 1983):

$$\text{True score } (\xi) = \sum_{1}^{n} [[(k_i+1)/k_i] \times p_i(\theta) - 1/k_i] ,$$

where n is the number of test items, and k_i+1 is the number of alternative answers of item i.

Equating true scores.

Suppose the ability level of an examinee on test for X is θ_x and ξ_x is the corresponding true score, and the ability level of the same examinee on alternate form Y is θ_y and ξ_y is the corresponding true score; then the equating equations for true scores are

$$\xi_x = \sum_{i=1}^{n} p_i(\theta_x) \text{ and } \xi_y = \sum_{j=1}^{m} p_j(\theta_y) \equiv \sum_{j=1}^{m} p_j(\alpha\theta_x + \beta) ,$$

where (1) n is the number of items on test X and m is the number of items on Y, (2) $p_i(\theta_x)$ is the probability of a correct answer to item i by an examinee, whose ability level on test X is θ_x, (3) $p_j(\theta_y)$ is the probability of a correct answer to item j by an examinee, whose ability level on test Y is θ_y, and (4) $\theta_y = \alpha\theta_x + \beta$ expresses the linear relationship between θ_y and θ_x (Hambleton & Swaminathan, 1990). In theory, for a given value θ_x, the pair of true scores (ξ_x, ξ_y) on tests X and Y can be determined. In practice, however, true scores can only be estimated.

Advantages of IRT Equating

Traditional equating methods can yield good results if the test forms are sufficiently parallel (Lord, 1980). However, when the tests to be equated differ in difficulties, IRT methods are considered to be better than linear methods. Major advantages of IRT equating include: (a) its flexibility in modeling either linear or curvilinear relationship between raw scores on alternate test forms, (b) equal reliability or identical observed score distributions is not assumed (Cook & Eignor, 1983; Kolen, 1981), (c) "item-free" estimates for persons and "person-free" item characteristics (Lord, 1977) are attainable, (d) unlike traditional equating methods, which only yield one single standard error of measurement for all examinees, error of measurement for ability estimation at each ability level can be estimated by IRT model, and (e) it may yield equivalent ability estimates for item sets differing in difficulty and/or discrimination, though not without measurement error (Green, Yen, & Burket, 1989).

Other appeals of IRT in practice are:

(1) It provides better equating at the upper end of the score scale, where important decisions are often made.

(2) It improves the flexibility in choosing among editions of a test, given the editions are placed on the same scale.

(3) If re-equating is needed, which usually occurs when certain items are added or dropped, it is easier to obtain the true score estimates with the IRT methods.

(4) It enables pre-equating, which derives the relationship between the test editions before they are administered operationally, given the pretest data are available (Cook & Eignor, 1983).

(5) For test forms across years that differ somewhat in content and length, the IRT equating may reduce the bias or scale drift in equating chains of circular-equating paradigm, and the stability of the scales near the extreme values may increase (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

Despite all the advantages listed above, Kolen and Brennan (1995) pointed out that IRT models gained their flexibility by making strong statistical assumptions and these assumptions were not likely to hold precisely in real testing situations. As a result, robustness of IRT models to the violations of model assumptions needs to be studied. Green, Yen, and Burket (1989) noted that it was not safe to say that IRT method would yield equivalent ability estimates if the items in different forms were different in content coverage. Therefore, test content should be carefully considered in IRT equating. Sometimes, the results of IRT equating agree with linear equating to a surprising degree. One possible explanation is that the test forms being equated are constructed to be similar considerably (Berk, 1982).

Effects of Examinee-Group Differences

Ideally, equating results should be independent of sub-populations of examinees of the same ability. Lawrence and Dorans (1990) suggested population independence be investigated under circumstances that examinee samples differed in ability.

Ability difference between examinee samples may have serious impacts on equating results (Cook, Eignor, & Schmitt, 1988). Theoretically, the closer the groups in ability, the more accurate the equating will be. However, Marco, Petersen, and Stewart (1983) found that if anchor test mirrored the content and difficulty level of the entire test, sample differences had relatively small and unsystematic effects on the quality of equating results.

Effect of Characteristics of Anchor items

The characteristics of anchor items, particularly the content representation and number of anchor items, may be influential on equating results.

Length of Anchor

Although there is no absolute standard for setting the number of an anchor items, a rule of thumb is to include at least 20 items or 20% of the total number of items in a test, whichever is larger (Angoff, 1984). Several studies have shown that as few as five or six carefully selected anchor items would perform satisfactorily for the IRT equating, when the item parameters of alternate tests were estimated by IRT concurrent method (Raju, Edwards, & Osberg, 1983; Wingersky & Lord, 1984; Raju, Bode, Larsen, & Steinhaus, 1988; Hills, Subhiyah, & Hirsch, 1988). Nevertheless, using IRT concurrent method, Hills, Subhiyah, and Hirsch (1988) found that randomly selected anchor items was not sufficient for producing satisfactory equating result, at least ten items was needed.

Content Representation

Whether anchor items are representative subset of the entire test, in terms of content and statistical properties, is especially important when examinee groups vary in ability (Cook & Petersen, 1987). Budescu (1985) pointed out that the magnitude of relationship between anchor test and unique components of each test form was the single most important determinant for the efficiency of equating. The relationship, however, depended on the reliability of the total test and the relative length of its two components. When non-random groups in an anchor design performed differentially, Budescu suggested that it was important to select anchor items that cover various content areas of a particular test to reflect the content mix of the entire test.

Equating Test Scores from Skewed Distributions

Often, equating is conducted for large scale achievement tests that have approximately symmetrical and bell-shaped score distributions. From time to time it is necessary, though, to equate tests that have skewed score distributions such as minimum-competency tests and licensure exams that have high passing standards. For licensure or certification programs, test forms are often equated with special interest on a particular cut-off score, or a range of scores, to inform decision making. To maximize the precision of the decision, it is reasonable to pay more attention to improve equating in the cutting score region, even at the expense of poorer equating at other scores (Brennan & Kolen, 1987).

Hills, Subhiyah, and Hirsch (1988) equated a minimum-competency test to an early version administered two years before and found that the results of the five equating methods used were generally similar to one another. They thus concluded that IRT equating methods could be applied to equating minimum-competency tests with extremely skewed distributions.

Assessing Equating Adequacy

Equating outcome can be evaluated in terms of its accuracy, sample invariance, and scale stability. This section of review focus on the estimation of equating accuracy, which is more relevant to the study design.

Criterion for Evaluating Equating Accuracy

It was found that IRT-based methods were better at equating both parallel and non-parallel tests (Kolen, 1981), effective for both inter-level and inter-form equating (Green, Yen, & Burket, 1989), and would yield more accurate equating outcomes than conventional equating (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988). These findings, however, may be tentative if the criterion used to evaluate the equating accuracy of IRT methods was biased. Therefore, in evaluating the effectiveness of various equating, it is important to seek a relatively unbiased criterion.

Often, equipercentile equating is used as evaluation criterion because it usually yields satisfactory results. In a comparative study, Livingston, Dorans, and Wright (1990) regarded equipercentile relationship as true equating relationship because true scores could be precisely estimated. Yen (1985) also suggested the use of equipercentile equating because it was as accurate as IRT equating.

Indices of Equating Accuracy

One common index used to represent equating accuracy is root-mean-squared deviation (RMSD), also known as root-mean-squared error of equating (RMSE). Suppose Form-B of a test is equated to Form-A, then

$$RMSD = \{[\sum_{y=1}^{v} n_y(\hat{x}_y - x_y)^2]/\sum n_y\}^{1/2},$$

where (a) n_y is the number of examinees with raw score y on Form-B, (b) $\hat{x}_y$ is the corresponding exact scaled score on Form-A determined by criterion equating, (c) x_y is the corresponding exact scaled score on Form-A determined by the equating to be evaluated against the criterion, and (d) the summation is over the raw-score levels on Form-B (Klein & Jarjoura, 1985; Livingston, Dorans, & Wright 1990).

Mean equating error, the bias that contributes to RMSD, can also be used as an index. It is estimated by: BIAS= $\bar{X} - \bar{X}'$, where $\bar{X}$ is the mean of the criterion scores and $\bar{X}'$ is the mean of the equivalents (Klein & Jarjoura, 1985). In addition, Marco, Petersen, and Stewart (1983) investigated the adequacy of curvilinear score equating by using squared bias and standardized weighted mean square difference, which weighted more on values that occurred more often, as indices of accuracy.

Dimensionality Issues

The robustness of IRT model to the violation of its assumptions is a major concern in IRT equating, because achievement tests usually cover multiple content topics, which may be influential on IRT model fit.

Definition of Unidimensionality

Test scores are most meaningful when all the items depend on a single trait. If the IRT assumption of unidimensionality holds, local independence should be observed. Statistically, local independence requires that, for fixed ability level θ, the item characteristic functions for any pair of items i and j should be independent (Lord, 1982b). If the probability for the given responses to the given items i and j are not independent at fixed θ, the responses may depend on some trait other than the θ. Hence, the IRT assumption of unidimensionality is violated.

Robustness of Unidimensionality Assumption

It has been shown that the violation of unidimensionality might have an impact on equating, but the effect might not be substantial (Dorans & Kingston, 1985). It depended on how the violation of the assumption is formulated. It was found that dimensionality violation would cause asymmetry of equating and influence the estimated magnitude of item discrimination parameter. However, similar equating outcomes were also found in equating tests differing in their dimensionality. It suggested that IRT equating might be robust against the violation of unidimensionality assumption. Or, it could be hypothesized that there was an overall ability, which could be conceptualized as a weighted composite of separate component abilities (Dorans & Kingston, 1985; Reckase, Ackerman, & Carlson, 1988; Yen, 1984).

Reckase, Ackerman, and Carlson (1988) had demonstrated that items measuring the same weighted composite abilities would meet the unidimensionality assumption for most of the IRT models. Dorans (1990) also argued that, although tests ought to measure the same construct and have the same content mix, they did not have to be composed of unidimensional items. If a test involved independent traits that influenced only a few items, Yen (1984) suggested that these traits might be ignored when the unidimensional trait was defined.

Limitations of Equating

Equating cannot solve problems originated in rough or improper test construction. It is mainly developed to improve on a test fairly constructed but fails to yield parallel forms. All conventional equating and IRT equating are primarily designed for test forms that have minor differences in their difficulties. Cook and Eignor (1991) indicated that no equating method could satisfactorily equate tests that were markedly different in difficulty, reliability or test content. As a result, there is a concern about the feasibility of vertical equating, which transforms scores across levels of achievement onto a single scale.

Due to floor and ceiling effects, tests that differ in difficulty are not likely to be equally reliable for all sub-groups of examinees (Skaggs & Lissitz, 1986). But, equal reliability is usually assumed in test equating such as linear equating and equipercentile equating. Thus it was argued that observed scores on tests differing in their difficulties cannot be equated. In practice, nevertheless, equating is conducted in its loose sense for a pragmatic purpose-- to approximate an ideal equivalency.

Description of Data

The particular test data used in this study has a rich content mixture (items were from 23 content sub-areas), which enables this study to investigate a variety of equating issues such as the characteristics of anchor items. Specifically, scores on the two forms, Book-A and Book-B, of a 1993 in-training examination taken by the candidates of a medical specialty were analyzed. The candidates took the test, while participating in various in-training programs located at different sites (usually in hospitals), to prepare for the board certification examination. No absolute score was used to determine pass or fail. The passing standard was 75% of the total test items being correctly answered.

To become board-certified, the candidates were strongly motivated to participate in the in-training programs for the preparation of the certification exams. Since the in-training test provided candidates valuable opportunities to get familiar with the formal certification exams, it was assumed that the candidates had taken the test as serious as when the formal exams were taken.

Test Content and Format

The test forms were comprised of five-alternative multiple-choice items, and the content of all the items were emergency-medicine-related. The item responses were all scored as right or wrong (coded as 1 or 0). Book-A had 203 items, of which 58 items were unique to Book-A. There were 52 unique items in Book-B, and the total number of items was 197. There were totally 145 anchor

items, and the anchors were identically embedded in both forms in terms of wording and location.

Examinee Groups

A total of 2,242 candidates took the in-training test. After screening the data, a case that had apparently guessed throughout the entire test was deleted from the analysis to secure the validity of scoring. Among the 2,241 subjects, 1,092 took Book-A and the rest of 1,149 took Book-B.

The examinee group taking Book-B scored higher in average on the anchor items, therefore it was likely that this group of examinees had higher ability. Nonetheless Lord (1981) mentioned, the difference in ability level would not influence equating result, given anchor-test design was employed. In addition, the group taking Book-B had a lower mean score on the unreduced full-length test. This implied that the unique items in Book-B had higher difficulty in average.

The test forms generally met the equating requirements that were mentioned earlier in the review of equating guidelines. Specifically, the test was reasonably long and all the items were from one single item pool. The anchor items constituted the major part of the total test. Some of the items were administered in the previous year under the same standardized testing situations. The size of the examinee groups, over 2,200 subjects, were reasonably large. In addition, the scoring key was clear and the test results appeared to be stable, given the preliminary analyses based on the classical test equating.

Research Design

Using four different item sampling schemes, pairs of test forms were assembled in this study with items sampled from the same big item pool. The various schemes were devised to manipulate the content mix of the tests, or the content representation of anchor items embedded in the test forms. The pairs of test forms were then equated, using anchor-item equating design with non-equivalent examinee groups, by Tucker linear method and two IRT methods. The equating results yielded by the different methods were compared against an appropriate criteria for evaluating equating accuracy that had several nice appeals.

Overall, content representation of anchor items and equating method are the two variables delineating the entire study. Other than the summary presented in Tables 1 and 2, basic research designs of this study are further elaborated in the following paragraphs.

Internal Anchor-Item Equating Design

The two examinee groups taking alternate test forms were not formed by random selection or assignment. Therefore, equating was made possible by the common items embedded in the alternate test forms. For the original test forms, the content of the anchor items was made representative to the entire test, and the anchor items were embedded in alternate test forms with same wording and at the same positions.

Manipulation of Content Representation of Anchor Items

All the items in the two original test forms are from a single big content domain. However, the content domain can be divided into 23 sub-content areas. Pooling together the items from the two original test forms, four subsets of items were drawn to form shorter test forms that had similar number of anchor items but the anchor items differed in their content representation. Thus the effect of content representativeness of anchor items on test equating could be studied. In general, the test lengths of all the shorter test forms (about 60 items) reflected the common test length seen in testing practice, and the various item sampling schemes used in this study were also used frequently in test construction.

Assumptions of Item Sampling Schemes

Various assumptions about the content of the test, used in this study, were made by the four item sampling schemes. They were briefly summarized in this section and details of the item sampling schemes and the sampling results were described in Appendix A.

Table 1

Summary of Basic Research Designs (1)

Item Sampling Scheme / Number of items / Alternate test form	Simple random sampling	Equal-weight domain random sampling	Proportional-weight domain random sampling	Purposeful sampling
Book-A	60 items	69 items	60 items	60 items
Book-B	60 items	69 items	60 items	57 items

Table 2

Summary of Basic Research Designs (2)

Criterion for Evaluating Equating Accuracy / Equating Method	Raw-score based criterion
IRT two-stage method	Index of accuracy-- Pearson r "pseudo true score", true score estimates
IRT fixed-b method	
Tucker linear method	

Note: Taking into account auto-correlation, for each test form, true score estimates were obtained by using (a) all of the items in the test form, (b) only the anchor items in the form, and (c) only the non-anchor items in the form.

The simple random sampling disregarded the existence of the 23 sub-content areas and randomly drew items from the big item pool. The equal-weight domain random sampling (random sampling stratified on sub-content areas) assumed that each of the 23 sub-content areas represented a significant part of the medical content domain, and these sub-content areas were equally important. The proportional domain random sampling assumed that the size of a sub-content area reflected its significance, therefore, it drew from each of the 23 sub-content areas a number of items proportional to the size of the area. And, the purposeful sampling included only the items from the largest three content sub-areas, assuming that the number of items in a sub-content area reflected the importance of the content.

If a test form involved a smaller number of sub-content areas, we would have more confidence in the assumption of unidimensionality made about the content of the test form.

Controlling for Anchor-length Effect

From a previous study using the same data, it was found that equating accuracy depended on the number of anchor items in the test forms being equated. Specifically, equating results from test forms that had longer anchor lengths tended to be more accurate (Yang & Houang, 1996). Therefore, in this study, the numbers of anchor items in various shorter test forms were fixed at 30 to avoid the confounding effect resulted from different anchor lengths. A number of 30 anchor items had been found to yield sufficiently accurate equating results.

Due to limited number of items available for item sampling, it was difficult to compose tests forms that all had the same number of anchor items. Nevertheless, this study ensured that at least 30 anchor items, a sufficient number of anchor items, were embedded in all of the shorter test forms.

Equating Methods

In addition to Tucker linear equating, two IRT-based methods were also used to equate alternate test forms for the study of method effect on equating accuracy. Both IRT-based equating are based on 3PL IRT model to account for guessing, because the chance for examinees to guess on some items could not be ruled out. One of the IRT-based method used is the two-stage method, which linearly transforms estimated IRT parameters on one test form to the parameter scales of another form. The second IRT-based method used is the fixed-b method, which sequentially calibrates test items of alternate forms. The two methods differed in their parameter-estimation procedures and, hence, equating procedures.

Criterion for Evaluating Equating Accuracy

In this study, items were sampled from one big item pool to form shorter test forms. As a result, examinee performances on the complete set of 145 common items in the big item pool could be regarded as the "anchor universe", relative to the anchor items embedded in the shorter test forms. "Pseudo true scores", the estimated true scores based on such "anchor universe", could be computed and thus used as eligible criteria for evaluating equating accuracy. However, such criterion was only appropriate when the examinee population and the testing occasion were considered fixed.

The "pseudo true score" was estimated by using the total raw score on the 145 anchor items. Although such raw-score-based criterion were susceptible to some drawbacks, including being person-dependent and item-dependent, it would not be biased in overestimating the accuracy of IRT equating. Intuitively, the lower bound of equating accuracy could be estimated for IRT equating. Therefore, the raw-score-based "pseudo true score" was chosen to represent a conservative criterion.

The accuracy of equating results were expressed by Pearson product moment correlation coefficient (r). A bigger positive Pearson r would indicate a more accurate equating result. Specifically, true scores based on various equating results from the shorter test forms were estimated and then correlated to "pseudo true scores" to obtain the indices of equating accuracy, the Pearson rs.

Research Tools

A variety of IRT calibration programs, such as ASCAL, BILOG, and LOGIST, were available for item and person estimation. The program chosen for the analyses of this study was the PC version BILOG. One advantage of using BILOG is that BILOG yields marginal maximum likelihood (MML) estimates and the number of parameters estimated does not increase with the increasing number of examinees. When the number of examinees increases, BILOG was found to yield more consistent results than LOGIST (Mislevy & Stocking, 1989; Baker, 1990). Yen (1987) also found that BILOG always yielded more precise estimates of individual item parameters. For shorter test with ten items, BILOG excelled LOGIST in estimating item and test characteristic functions; whereas for longer tests with 20 to 40 items, the two programs yielded similar estimates. Mislevy and Stocking (1989) also found that BILOG would yield more reasonable results when the examinee samples are smaller.

In addition to BILOG, SAS for Unix and Excel spreadsheet were also used in this study to facilitate Tucker linear equating and all other sorts of data management and analyses.

Research Limitations

The scope and depth of this study was limited by personal interest and ability. Environmental conditions, such as the cost, the availability, and the capacity of computer packages for IRT calibration and equipercentile equating, also set limits for this study. Despite the fact that the rich context of the data analyzed in this study helped enrich the research questions and the study design, the data analyzed still set limits for this study in the sense that it was secondary data so any manipulations before and during data collection were not accessible. For instance, equating using anchor-item design was the only option for this study because the test forms were written with embedded anchor items and given to non-equivalent groups.

Results and Discussion

Results of classical item analyses, correlation analyses on anchor items and none-anchor items, inspection on examinee group differences, IRT parameter estimations, equating outcomes yielded by various equating methods, as well as the evaluation of equating accuracy are all presented and discussed in this section. Issues concerning the use of the index of equating accuracy, the adequacy of 3PL IRT model, as well as the validity and reliability of anchor items are also considered.

Classical Item Analyses

Analyses on item difficulties showed that in general average item difficulties, ranging from 0.688 to 0.759, were quite similar for the four pairs of test forms and were considered moderate. The standard deviations of item difficulties within various test forms were also very similar, ranging from 0.145 to 0.153. These small standard deviations implied that items within the same test forms generally did not differ much in their difficulties. The distribution plots shown in Figure 1 further indicated that item difficulties were evenly spread within test forms for all pairs of test forms. Distributions of item-total correlation were presented in Figure 2. It was found that item scores generally correlated moderately to total test scores for all the test forms.

In summary, classical item analyses suggested that (a) the alternate test forms created in this study did not differ much in item difficulty and item-total correlation, thus were good candidates for equating, and (b) the four pairs of test forms looked quite similar to one another in terms of average difficulty, which provided a fair basis for the study of the effect of anchor characteristics on equating accuracy.

Representation of Anchor Items

Results of correlation analyses on anchor items and none-anchor items (see Appendix B) provided a closer look at the composition of various test

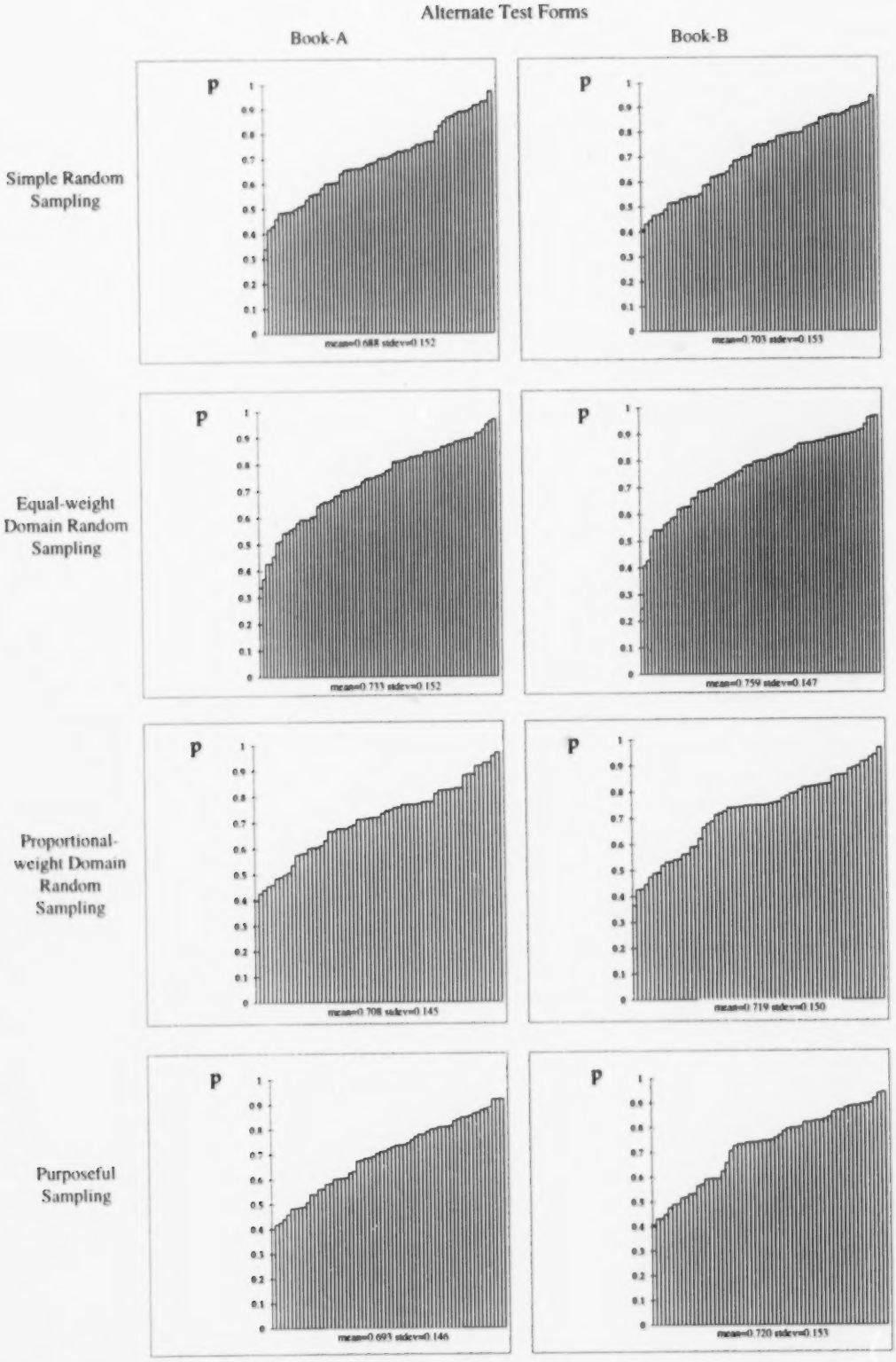

Figure 1. Distributions of item difficulty (p) for pairs of shorter test forms

fig1.2 wk 3/21/97

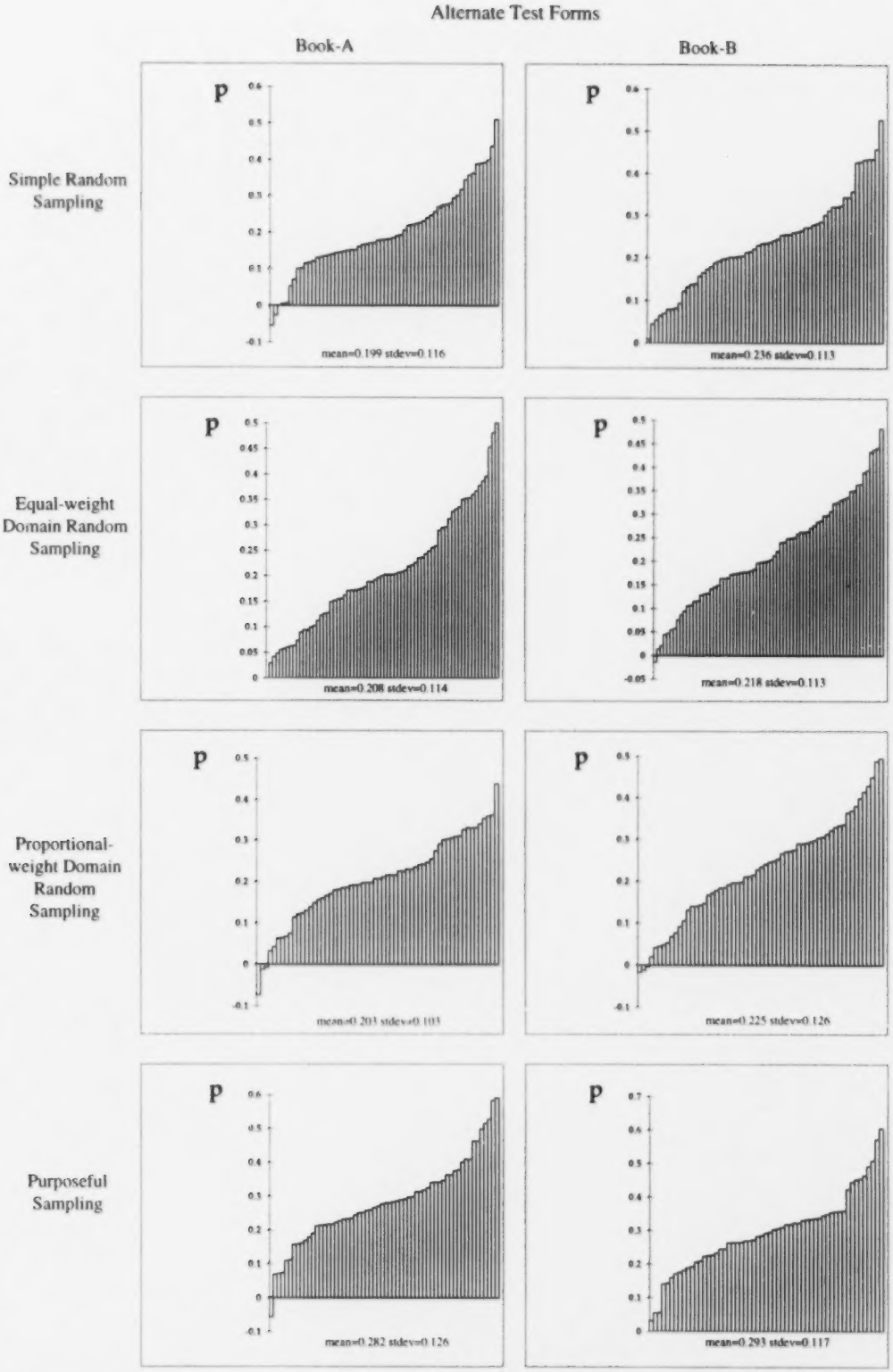

Figure 2. Distributions of item-total correlation for pairs of shorter test forms

forms. Overall, for each of the shorter test forms and the original longer test forms, anchor items and none-anchor items correlated to each other significantly to a moderate degree. The correlation coefficients ranged from .44 to .54 across various shorter test forms. Both anchor and none-anchor items of a test also correlated significantly with the entire test to a considerable degree. The correlation between anchor items and the entire test ranged from .86 to .97 across various test forms. As a result, it seemed reasonable to use the anchor items to equate the entire alternate forms.

A shrinking trend was found in the correlation between anchor items and entire test across various test forms. In summary, (a) for test form Book-A, the magnitude of correlation coefficient decreased from .97 of purposeful sampling, to .94 of equal-weight domain random sampling, to .92 of proportional-weight domain random sampling, and to .86 of simple random sampling; and (b) for Book-B, the pattern of shrinkage remained, and the coefficient dropped from .97 to .94 to .93 to .86 accordingly. The shrinkage suggested that content representation of anchor items was likely to very with item sampling schemes. The purposeful sampling seemed to have yielded anchor items that were most representative of the entire test. It made sense because all of the items sampled by this scheme concentrated on merely three sub-content areas and were likely to be more similar in content. The simple random sampling scheme resulted in anchor items that seemed least representative. The finding could be attributed to the fact that the randomly sampled items scattered all over 23 sub-content areas such that the overall content was more heterogeneous. The similar results of equal-weight and proportional-weight domain random sampling might reflect the indifference between sampling items evenly from all sub-content areas and having more emphasis on larger sub-content areas.

It should be noted, though , the magnitude of the correlation between anchor items and the entire test was inflated by auto-correlation because the internal anchor was a subset of the test. The magnitude of auto-correlation depended greatly on the number of anchor items embedded in a test. Consequently, whether anchor items were representative of the entire test should not be solely determined by looking at the correlation coefficient. In this study, however, the effect of auto-correlation were expected to be about the same on various test forms because their anchor lengths were fixed to be similar.

Considerations of Group Differences

Overall, the average raw scores of examinee groups taking different test forms did not differ substantially. Upon a closer inspection on the raw scores, however, it was found that examinees taking one test form (Book-B) scored slightly higher than examinees taking the other form (Book-A) on both anchor items and unique items across all pairs of shorter test forms.

To further inspect examinee group differences, the average item difficulties broken down by test form and type of items were computed for all of the test forms. The results of the average item difficulties were summarized in Table 3. Slightly larger percentages were found consistently over various test forms for examinees taking Book-B on anchor items, indicating that the examinees might have higher ability than examinees taking Book-A. The group differences were probably due to the non-random selection or assignment of examinees in testing.

As discussed previously in literature review, examinee-group disparity may be a threat to the equating accuracy of Tucker linear method, therefore Levine equally reliable method is sometimes recommended instead (Kolen & Brennan, 1987). In this study, however, Tucker method was still used because (a) the differences found between examinee groups were small and equating results of Tucker method were expected not to be affected, (b) the advantage of Levine method over Tucker method is still not clearly known (Kolen & Brennan, 1987), (c) Levine method generally is more appropriate for more similar test forms, but the similarity between the test forms used in this study was not clearly confirmed, and (d) it was found that equating results yielded by the two methods for the original test forms were almost identical,

Table 3

Average Item Difficulty ($\overline{p}$) by Item Type by Test Form

		Simple Random Sample	Equal-weight Domain Random Sample	Proportional-weight Domain Random Sample	Purposeful Sample
Anchor Items	Book-A	0.722	0.767	0.720	0.703
	Book-B	0.736	0.781	0.734	0.728
Unique Items	Book-A	0.654	0.650	0.683	0.664
	Book-B	0.670	0.705	0.690	0.691

Note: 1. Let P_i= the % of examinees getting item "i" right,

then $\overline{p} = \sum_{i=1}^{n} p_i$; n = total # of items.

2. In this study, 1,092 examinees took Book-A, and 1,149 took Book-B.

thus it was safe to conclude that the two methods would make no difference for the test analyzed in this study.

Estimation of IRT Parameters

Results of IRT item and person parameter estimations were summarized in Table 4 for the four pairs of test forms. Roughly, the patterns of estimated parameters showed that test forms created by different item sampling schemes differed less in their average item discrimination but more in their item difficulties. The mean item difficulties on anchor items also differed across test forms, and the mean item difficulties for the test forms created by purposeful sampling of item looked especially different from the rest. The differences in the estimated item difficulties seemed suggesting some effect of item sampling on test and anchor characteristics. Comparing the mean item difficulties on anchor items for the two alternate forms (Book-A and Book-B), it should be noted that purposeful sampling seemed to have created test forms that were more different than the forms created by the other item sampling schemes.

Equating Ability Estimates

Equated IRT ability estimates yielded by two-stage and fixed-b methods were correlated to compare the equating results of the two methods. Pearson correlation coefficients were computed and the results for various test forms are as follows: (a) r=0.99985 for the test form composed by simple random sampling of items, (b) r=0.99961 for equal-weight domain random sample, (c) r=0.99961 for proportional-weight domain random sample, and (d) r=0.99993 for purposeful sample. These nearly perfect and significant correlation strongly suggested that the two IRT equating methods were almost identical in determining the standings of individual examinees in a group. It could be argued that there was no IRT method effect on ability estimation in this study.

Estimation of True Scores

To obtain true score estimates, the following formula was used (Lord, 1980):

$$\text{Estimated true score } (\hat{T}) = \sum_{i=1}^{n} p_i(\theta) = \sum_{i=1}^{n} \{c_i + (1-c_i)/[1 + Exp^{-1.7a_i(\theta-b_i)}]\} \ ,$$

where θ is examinee ability and n is the number of items.

As expected, for all test forms, the correlation between estimated true scores based on the two IRT equating was almost perfect and significant. It was consistent with the findings on the IRT estimated ability estimates. Thus it was concluded that the two IRT equating methods were not different in equating the tests in this study and would place individual examinees of a group in almost the same order.

Results of Tucker Linear Equating

For each pair of alternate test forms, Tucker linear method was applied to find an equating equation for transforming scores on Book-B to a set of new scores comparable to scores on Book-A. The Tucker equating equations derived for the four shorter test were presented in Table 5, along with a summarization of important statistics used to arrive at the equations. Using the Tucker equations, equivalent scores were established for test forms Book-A and Book-B.

Evaluation of Equating Accuracy

The total raw scores of examinees on all the 145 common items in the original item pool were computed and treated as the "pseudo true scores". The "pseudo true scores" were then correlate with the estimated IRT true scores yielded by the two IRT equating, as well as the scaled total scores obtained by Tucker linear method. Pearson correlation coefficients were computed and used as indices of equating accuracy. Specifically, a positive and bigger coefficient would indicate a more accurate equating result. The collection of correlation coefficients between the "pseudo true scores" and the estimated true score yielded by various equating method for various test forms were presented in the big correlation matrix in Table 6 to illustrate the accuracy

Table 4
Results of IRT Parameter Estimation

Alternate forms		Estimated parameter		Simple random sampling	Equal-weight domain random sampling	Proportional-weight domain random sampling	Purposeful sampling
Book-A		$\hat{a}$	mean	0.340	0.342	0.340	0.444
			s.d.	0.173	0.168	0.127	0.192
		$\hat{b}$	mean	-0.884	-1.445	-1.090	-0.653
			s.d.	2.239	2.231	2.043	1.848
		$\hat{c}$	mean	0.252	0.260	0.256	0.247
			s.d.	0.046	0.033	0.029	0.052
		$\bar{b}_{anchor}$		-1.340	-1.750	-1.090	-0.750
		$\hat{\theta}$	mean	0.003	0.006	0.005	0.007
			s.d.	0.851	0.854	0.839	0.897
Book-B	Using IRT two-stage method	$\hat{a}$	mean	0.377	0.355	0.410	0.444
			s.d.	0.165	0.162	0.162	0.041
		$\hat{b}$	mean	-1.008	-1.561	-0.380	-0.904
			s.d.	1.891	2.240	2.361	1.705
		$\hat{c}$	mean	0.241	0.270	0.328	0.231
			s.d.	0.034	0.030	0.052	0.041
		$\bar{b}_{anchor}$		-1.45	-1.88	-0.840	-0.180
		$\hat{\theta}$	mean	0.003	0.004	0.012	0.005
			s.d.	0.868	0.857	0.858	0.886
	Using IRT fixed-b method	$\hat{a}$	mean	0.400	0.377	0.433	0.462
			s.d.	0.164	0.165	0.166	0.194
		$\hat{b}$	mean	-0.591	-1.200	0.052	-0.650
			s.d.	1.951	2.374	2.301	1.774
		$\hat{c}$	mean	0.311	0.347	0.384	0.277
			s.d.	0.053	0.048	0.056	0.049
		$\hat{\theta}$	mean	0.059	0.011	0.142	0.061
			s.d.	0.880	0.867	0.869	0.888

Note: a = item discrimination parameter
b = item difficulty parameter
c = guessing parameter
θ = person ability parameter
$\bar{b}_{anchor}$ =mean anchor item difficulty

BEST COPY AVAILABLE

Table 5
Summary of the Results of Tucker Linear Equating

Alternate Test Forms / Estimates / Test Assembling Method	Book-A			Book-B			Tucker Equating Equation
	$\alpha_A(A\|V)$	$\mu_s(A)$	$\sigma_s^2(A)$	$\alpha_B(B\|V)$	$\mu_s(B)$	$\sigma_s^2(B)$	
Simple random sampling	1.524	41.615	32.742	1.593	41.855	36.047	$\ell(b)$=.953(b-41.855)+41.615
Equal-weight domain random sampling	1.218	51.003	33.830	1.257	51.941	35.559	$\ell(b)$=.975(b-51.941)+51.003
Proportional-weight domain random sampling	1.257	42.810	30.935	1.291	42.806	32.349	$\ell(b)$=.978(b-42.807)+42.810
Purposeful sampling	1.169	42.271	44.882	1.150	40.408	43.036	$\ell(b)$=1.021(b-40.408)+42.271

Note:

1. "A" represents test form "Book-A", "B" represents test form "Book-B", and "V" represents common items.
2. α_A is the regression coefficient for the population taking Book-A, and α_B is the regression coefficient for the population taking Book-B.
3. "*s*" denotes the synthetic population, and "*b*" is the observed score on Book-B.
4. The weight for the population taking Book-A is .487, and the weight for the population taking Book-B is .513.

Table 6

Correlation Matrix for Evaluating Equating Accuracy

(Index of accuracy-- Pearson r between 'pseudo true score' and true score estimate)

		"Pseudo True Score"	Tucker Linear Method				IRT Two-stage Method				IRT Fixed-b Method			
			Equal-weight Domain Random Sampling	Purposeful Sampling	Proportional-weight Domain Random Sampling	Simple Random Sampling	Equal-weight Domain Random Sampling	Purposeful Sampling	Proportional-weight Domain Random Sampling	Simple Random Sampling	Equal-weight Domain Random Sampling	Purposeful Sampling	Proportional-weight Domain Random Sampling	Simple Random Sampling
"Pseudo True Score"		1.000												
Tucker Linear Method	Equal-weight Domain Random Sampling	0.859	1.000											
	Purposeful Sampling	0.892	0.755	1.000										
	Proportional-weight Domain Random Sampling	0.860	0.782	0.795	1.000									
	Simple Random Sampling	0.832	0.782	0.810	0.795	1.000								
IRT Two-stage Method	Equal-weight Domain Random Sampling	0.877	0.964	0.776	0.787	0.786	1.000							
	Purposeful Sampling	0.894	0.739	0.973	0.771	0.795	0.771	1.000						
	Proportional-weight Domain Random Sampling	0.845	0.750	0.773	0.944	0.768	0.776	0.748	1.000					
	Simple Random Sampling	0.856	0.780	0.834	0.791	0.965	0.811	0.842	0.785	1.000				
IRT Fixed-b Method	Equal-weight Domain Random Sampling	0.873	0.961	0.770	0.785	0.783	**0.997**	0.759	0.789	0.808	1.000			
	Purposeful Sampling	0.895	0.740	0.972	0.774	0.799	0.772	**0.989**	0.782	0.846	0.771	1.000		
	Proportional-weight Domain Random Sampling	0.870	0.771	0.801	0.963	0.785	0.798	0.793	**0.976**	0.804	0.795	0.795	1.000	
	Simple Random Sampling	0.854	0.779	0.831	0.790	0.963	0.810	0.837	0.791	**0.999**	0.810	0.846	0.802	1.000

Note. All of the Pearson correlation coefficients are significant at α=.01.

BEST COPY AVAILABLE

of various equating.

Comparisons Among Equating Methods

Overall, the indices of accuracy (Pearson correlation coefficients) ranged from .832 to .894 across various test forms. It seemed that the equating results yielded by the three equating methods were all accurate to a moderate degree, and examinees were generally ordered in a consistent way, no matter which method was used.

Despite the fact that the indices of accuracy in Table 6 all looked similar, IRT equating appeared to have yielded more accurate results than Tucker linear method always. The only exception occurred when the test forms composed by proportional-weight domain random sampling were equated, where Tucker method (r=.860) seemed to do better than IRT two-stage method (r=.845). The results of the two IRT methods correlated strongly and the rs ranged from .976 to .999 (see the bolded numbers in Table 6), showing that the IRT methods yielded very similar results. The results of Tucker method, however, correlated less strongly to the IRT results, with rs ranging from .944 to .973 (see the underscored numbers in Table 6).

Comparisons Among Test Forms

Comparing the equating results on various test forms, it was found that both Tucker and IRT methods worked best for the forms composed by purposeful item sampling scheme, where the index of accuracy was .895 in average. The methods seemed to yield the least satisfactory results for the forms based on simple random sampling of items, where the mean accuracy was .847. In addition, the average accuracy for the test forms based on proportional-weight and equal-weight random sampling were .858 and .869 respectively, indicating a similarity in the item sampling effects of the two schemes.

Effect of Content Representation of Anchor Items

As discussed earlier, purposeful sampling yielded the most representative anchor items and random sampling resulted in anchor items that were least representative of the entire test. Combined the findings with the above outcomes, it seemed reasonable to conclude that equating accuracy might depend on the content mix or the content representativeness of anchor items. That is, Tucker linear method and the two IRT methods are more likely to yield more accurate results when anchor items are more similar to the entire test, or the content coverage of a test concentrates on fewer topics. In short, the characteristics of anchor items may have substantial impacts on the accuracy of test equating, regardless of the equating method used. As a result, to improve equating accuracy, it is important to include anchor items that can fully reflect overall content coverage of the entire test.

Controlling Artifact due to Auto-correlation

For the index of accuracy, there was a concern about auto-correlation caused by the fact that the "pseudo true score" was computed based on the complete set of 145 anchor items and the anchor items in shorter test forms were part of the complete anchor set. Due to the overlapping of items, correlation coefficients that showed the relationship between true scores and estimated true scores were inflated. To unmask the relationship to better estimate equating accuracy, "pseudo true scores" were correlated with the estimated IRT true scores that involved none-anchor items only. The results of correlation analyses were summarized in Appendix C. The same strategy for controlling auto-correlation, however, was not applied to Tucker linear equating. Because Tucker method is based on observed test score as a whole, unlike IRT methods that are more flexible in calibrating revised tests, it is not feasible to obtain scaled scores on non-anchor items only.

After controlling the artifact due to auto-correlation, the patterns of rs found among various methods and test forms in previous section remain unchanged. The problem of auto-correlation seemed not to be serious, therefore the conclusions about the accuracy of various equating methods on different test forms and the effect of anchor characteristics were retained.

Although the threat from auto-correlation may not be completely eliminated by removing anchor items from the correlation analyses, by controlling part of the artifact, the set of new indices of accuracy would

provide a better opportunity for understanding the effectiveness of equating methods.

Concurrent Validity and Reliability of Anchor Items

The data was further exploited to investigate the validity and reliability of anchor items, by correlating "pseudo true scores" with IRT estimated true scores using anchor items only. The results were summarized in Appendix D. Because the anchor items included in shorter test forms were part of the set of 145 anchor items, from which "pseudo true score" was derived, "pseudo true score" could also be regarded as a similar but more reliable measure for the anchor items. From this perspective, "pseudo true score" was used as a criterion measure to study the concurrent validity of anchor items, and Pearson correlation coefficient was computed as a measure of validity. Furthermore, by correlating an observed score (the estimated true score) with its corresponding true score (the "pseudo true score"), the correlation coefficient may be regarded as a reliability measures for the observed score. From this point of view, the Pearson rs in Appendix D were also measures of reliability.

In summary, strong relationship was found between "pseudo true score" and anchor items for each of the shorter test. It provided some evidence of validity and reliability for anchor items. In average, the validity/reliability coefficient was .894 for the anchor items of the test form composed by purposeful item sampling schemes, .875 and .858 for the anchor items sampled by equal-weight and proportional-weight schemes, and .856 for the anchor items drawn by simple random sampling. Given the validity and reliability evidence for anchor items used for equating, along with the equating accuracy found, both IRT equating methods were concluded to be satisfactory.

Limitation of the Criterion for Evaluating Equating Accuracy

As described earlier in the section of research design, "pseudo true score", the raw-score-based criterion for evaluating equating accuracy, is conceptually reasonable and will not over-estimate the accuracy of IRT equating. The evidence of reliability and validity, as well as the availability of data from all examinees, also support the use of the criterion. Nonetheless, it is limited in the following senses: (a) it is only appropriate when examinee group and testing occasion are considered fixed, as noted earlier, (b) in essence, it remains a convenient close estimate of true score that has measurement error, and (c) it is susceptible to problems such as person-depend and item-dependent, due to its raw-score-based nature.

Alternatively, IRT estimated score can be computed using the 145 common items and used as another type of "pseudo true score" or criterion for evaluating equating accuracy. However, it is known that such IRT-based criterion may be biased in over-estimating the accuracy of IRT equating, while underestimating the accuracy of linear equating. Taking into account all the facts, the raw-score-based criterion was used in this study because it would provide a conservative estimate of equating accuracy for IRT equating.

Adequacy of 3PL IRT Model

The results of using the item and person parameter estimates of 3PL IRT model for equating the minimum competence test analyzed in this study seemed adequate. As explained earlier, the use of 3PL IRT model for parameter estimation is a logical choice. In addition, the satisfactory equating results yielded by the two IRT equating methods also help justify its use. It can thus be concluded that it is appropriate to include guessing parameter when tests or test forms with negatively skewed score distributions are equated.

Suggestions

Equating accuracy can be better estimated if unbiased evaluation criteria are identified and used. To compensate for the arbitrary and often biased nature of common criteria used for evaluating equating accuracy, multiple criteria can be devised to estimate equating accuracy so the estimation outcomes can be compared to determine the relative effectiveness of

these criteria. Therefore, in a subsequent study, several other criteria are proposed to evaluate equating accuracy, including a different "pseudo true score" based on estimated IRT true score using the 145 anchor items, and the results of equipercentile equating, which are often considered satisfactory.

Assuming unidimensionality, in this study, 3PL IRT model seemed to have yielded satisfactory estimates that were used to derive equivalent scores in subsequent equating process. However, due to the fact that there are 23 sub-content areas nested within the big content domain for the test, whether the assumption of unidimensionality holds seems ambiguous.
If there are in fact more than one underlying traits for the test, then the findings of this study suggest that the IRT model used is robust to the violation of unidimensionality assumption. Nevertheless, in such case, multidimensional IRT models may yield better results than the unidimensional model. Therefore, dimensionality of the test should be carefully inspected or defined via theoretical review, content analysis, or factor analysis so IRT item and person parameters can be better estimated and used in equating.

For some other minimum competency test, if guessing effect is considered not serious, then the use of Rasch model or 2PL IRT model may be better alternatives to the 3PL IRT model. More investigations are needed for the data-model fit of IRT parameter estimation, since the estimation results may have substantial impacts on equating accuracy.

Beyond the current study, it will be intriguing to investigate functions of various equating methods when test forms become longer or the number of anchor items increases. Cross-year equating can also be conducted to examine the effects of equating over time. If possible, validation study can also be carried out to further determine equating accuracy by correlating equating outcomes to the testing outcomes of some other examinations that need no equating.

Reference

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. Applied Psychological Measurement, 14, 139-150.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. Journal of Educational Measurement, 28, 147-162.

Berk, R. H. (1982). Discussion of item response theory. In P. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Berry, D. A., & Lindgren, B. W. (1990). Statistics: theory and methods Belmont, CA: Brooks/Cole.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland and D. B. Rubin (Eds.) Test equating (pp. 9-49). New York: Academic.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. Applied Psychological Measurement, 11, 279-290.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22, 13-20.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory (pp.175-195). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. Educational Measurement: Issues and Practice, 10, 37-45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988). The effects on IRT and conventional achievement test equating results of using equating samples matched on ability (Research Rep. No. RR-88-52). Princeton, NJ: Educational Testing Service.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Chicago: Holt, Rinehart and Winston, Inc.

Dorans, N. J. (1990). Equating methods and sampling designs. Applied Measurement in Education, 3, 3-17.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of Educational Measurement, 22, 249-262.

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2, 297-312.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. Japanese Psychological Research, 22, 144-49.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R. K., & Swaminathan, H. (1990). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. Journal of Educational Measurement, 25, 221-231.

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions (Technical Report No. 87-79). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1989). The kernel method of equating score distributions (Technical Report No. 89-84). Princeton, NJ: Edcuational Testing Service.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. Journal of educational measurement, 22, 197-206.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18, 1-10.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. Applied Psychological Measurement, 11, 263-277.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Kolen, M. J., & Jarjoura, D. (1987). Analytical smoothing for equipercentile equating under the common item nonequivalent populations design. Psychometrika, 52, 43-59.

Kaolin, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. Journal of Educational Measurement, 27, 27-39.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. Applied Measurement in Education, 3, 19-36.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? Applied Measurement in Education, 3, 73-95.

Lord, F. M. (1965). A strong true score theory with applications. Psychometrika, 30, 239-270.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1982a). The standard error of equipercentile equating. Journal of Educational Statistics, 1, 165-192.

Lord, F. M. (1982b). Item response theory and equating- A technical summary. In P. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch Model. Journal of Educational Measurement, 17, 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 14, 139-160.

Marco, G. L., Petersen, N. C., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), New horizons in testing: Latent trait theory and computerized adaptive testing. New York: Academic Press.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational Measurement. New York: ACE/Macmillan.

Raju, N. S., Bode, R. K., Larsen, V. S., & Steinhaus, S. (1986, April). Anchor-test size and horizontal equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983, April). The effect of anchor test size in vertical equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, 25, 193-203.

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. British Journal of Mathematical and Statistical Psychology, 40, 43-49.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Wingersky, M. S., & Barton, M. A. (1982). Logist user's guide: Logist 5, Version 1.0. Princeton, NJ: Educational Testing Service.

Yang, W. L., & Houang, R. T. (1996, April). The effect of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT-based equating using anchor-item design. Paper presented at the annual meeting of the American Educational Research Association, New York.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1985). Tau equivalence of vertical equating using three-parameter item response theory and Thurstonian procedures. Paper presented at the meeting of the American Educational Research Association, Chicago.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items. Chicago: Scientific Software International.

Appendix A
Item Sampling Schemes for Shorter Test Forms

Simple Random Sampling

Assumption

Items from different sub-content areas do not differ substantially, since all of them are written for one single content domain (medicine-related) .

Method

Pool and mix items from all of the 23 sub-content areas to form a big item pool. Then, randomly sample items from the pool using a random number table.

Results

One pair of shorter alternate test forms, each consisting 60 items. There are 30 anchor items in each test form.

Equal-Weight Domain Random Sampling

Assumption

Each of the 23 sub-content areas represents an important part of the medical content domain, and the 23 areas are of equal importance.

Method

For the first test form, sample three items from each of the 23 sub-content areas, regardless of the size of these areas. To have anchor items spread evenly across various areas and to account for the fact that there are more anchor items in the big item pool, whenever it is possible, two anchor items and one none-anchor item are randomly drawn from each of the areas. Use the anchor items sampled for the first test form as the anchor items of the second test form, and randomly sample one none-anchor item from each of the content areas to make up the entire second test form.

Result

A pair of alternate test forms, each consisting 69 items. For each test form, there are 49 anchor items and 20 none-anchor items.

Proportional-weight Domain Random Sampling

Assumption

The size of a sub-content area reflects its importance, that is, the more items a sub-content area has, the more important the area is to the medical content domain.

Method

From each of the 23 sub-content areas, randomly sample a number of items that is proportional to the size of the sub-content area. The sampling procedure is illustrated below in more details:

Content Area	Area size (total # of items)	%	# of items to be sampled
1	13	5.8	5.8 * 60 = 3.48 ≅ 4
2	23	10.2	10.2 * 60 = 6.12 ≅6

3	3	1.3	1.3 * 60 = 0.78 ≅1
4	14	6.2	6.2 * 60 = 3.72 ≅4
5	5	2.2	2.2 * 60 = 1.32 ≅1
6	19	8.4	8.4 * 60 = 5.04 ≅5
7	5	2.2	2.2 * 60 = 1.32 ≅1
8	3	1.3	1.3 * 60 = 0.78 ≅1
9	6	2.7	2.7 * 60 = 1.62 ≅2
10	9	4.0	4.0 * 60 = 2.40 ≅2
11	8	3.6	3.6 * 60 = 2.16 ≅2
12	4	1.8	1.8 * 60 = 1.08 ≅1
13	13	5.8	5.8 * 60 = 3.48 ≅4
14	7	3.1	3.1 * 60 = 1.86 ≅2
15	5	2.2	2.2 * 60 = 1.32 ≅1
16	15	6.7	6.7 * 60 = 4.02 ≅4
17	13	5.8	5.8 * 60 = 3.48 ≅4
18	25	11.1	11.1 * 60 = 6.66 ≅7
19	8	3.6	3.6 * 60 = 2.16 ≅2
20	5	2.2	2.2 * 60 = 1.32 ≅1
21	4	1.8	1.8 * 60 = 1.08 ≅1
22	9	4.0	4.0 * 60 = 2.40 ≅2
23	9	4.0	4.0 * 60 = 2.40 ≅2
Total	225	100.0	60

Result

A pair of alternate test forms, each consisting 60 items. In each form, there are 40 anchor items.

Purposeful Sampling

Assumption

The more items a sub-content area has, the more important the area is, and the 23 sub-content areas differ in their content to a somewhat degree. In other words, test form involving a smaller number of content areas will be more homogeneous in content.

Method

Include all of the items in the largest three content areas, and disregard any items in the rest of the areas.

Result

For one test form, 45 anchor items and 15 none-anchor items are included. And, for the other test form, there are 45 anchor items and 12 none-anchor items.

Appendix B

Correlation Analyses on Anchor and None-anchor Items

Correlation Coefficient / Test		$r_{(anchor,non\text{-}anchor)}$		$r_{(anchor,entire\ test)}$		$r_{(non\text{-}anchor,entire\ test)}$	
		Alternate Forms: Book-A	Book-B	Book-A	Book-B	Book-A	Book-B
Original test		.754**	.736**	.981**	.979**	.866**	.859**
Shorter test by …	Simple random sampling	.503**	.537**	.861**	.863**	.873**	.889**
	Equal-weight domain random sampling	.439**	.488**	.939**	.939**	.721**	.758**
	Proportional-weight domain random sampling	.443**	.482**	.924**	.925**	.752**	.778**
	Purposeful sampling	.486**	.451**	.968**	.968**	.690**	.660**

Note: *- Signif. LE 0.05 **- Signif. LE 0.01 (2-tailed)

BEST COPY AVAILABLE

Appendix C

Correlation Matrix for Evaluating Equating Accuracy, with a Control of Auto-Correlation

(Index of accuracy-- Pearson r between 'pseudo true score' and true score estimate for none-anchor items only)

		"Pseudo True Score"	IRT Two-stage Method				IRT Fixed-b Method			
			Equal-weight Domain Random Sampling	Purposeful Sampling	Proportional-weight Domain Random Sampling	Simple Random Sampling	Equal-weight Domain Random Sampling	Purposeful Sampling	Proportional-weight Domain Random Sampling	Simple Random Sampling
"Pseudo True Score"		1.000								
IRT Two-stage Method	Equal-weight Domain Random Sampling	0.868	1.000							
	Purposeful Sampling	0.891	0.764	1.000						
	Proportional-weight Domain Random Sampling	0.842	0.771	0.749	1.000					
	Simple Random Sampling	0.854	0.809	0.843	0.784	1.000				
IRT Fixed-b Method	Equal-weight Domain Random Sampling	0.865	**0.997**	0.752	0.784	0.806	1.000			
	Purposeful Sampling	0.892	0.765	**0.989**	0.783	0.847	0.764	1.000		
	Proportional-weight Domain Random Sampling	0.867	0.793	0.794	**0.975**	0.803	0.790	0.795	1.000	
	Simple Random Sampling	0.852	0.808	0.838	0.790	**0.999**	0.807	0.847	0.802	1.000

Note: All of the Pearson correlation coefficients are significant at $\alpha = .01$.

BEST COPY AVAILABLE

Appendix D

Correlation Matrix for Reliability and Validity of Anchor Items

(Pearson r between 'pseudo true score' and true score estimate using anchor items only)

		"Pseudo True Score"	IRT Two-stage Method				IRT Fixed-b Method			
			Equal-weight Domain Random Sampling	Purposeful Sampling	Proportional-weight Domain Random Sampling	Simple Random Sampling	Equal-weight Domain Random Sampling	Purposeful Sampling	Proportional-weight Domain Random Sampling	Simple Random Sampling
"Pseudo True Score"		1.000								
IRT Two-stage Method	Equal-weight Domain Random Sampling	0.877	1.000							
	Purposeful Sampling	0.894	0.772	1.000						
	Proportional-weight Domain Random Sampling	0.846	0.777	0.748	1.000					
	Simple Random Sampling	0.857	0.811	0.842	0.786	1.000				
IRT Fixed-b Method	Equal-weight Domain Random Sampling	0.873	0.997	0.760	0.790	0.808	1.000			
	Purposeful Sampling	0.895	0.772	0.989	0.782	0.846	0.771	1.000		
	Proportional-weight Domain Random Sampling	0.871	0.798	0.793	0.976	0.804	0.796	0.795	1.000	
	Simple Random Sampling	0.855	0.810	0.837	0.792	0.999	0.810	0.846	0.803	1.000

Note: All of the Pearson correlation coefficients are significant at α=.01